Prior Based Pyramid Residual Clique Network for Human Body Image Super-Resolution

Simiao Wang[a,b], Yu Sang[b,], Yunan Liu[a,], Chunpeng Wang[c], Mingyu Lu[a], Jinguang Sun[b]

[a]*School of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China*
[b]*School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China*
[c]*Shandong Provincial Key Lab of Computer Networks, School of Cyber Security,*
Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China

Abstract

Recent research in the analysis of human images, such as human parsing and pose estimation, usually requires input images to have a sufficiently high resolution. However, small images of people are commonly encountered in our daily lives, particularly in surveillance applications. This paper aims to ultra-resolve a tiny person image to its high-resolution counterpart by learning effective feature representations and exploiting useful human body prior knowledge. First, we propose the Residual Clique Block (RCB) to fully exploit compact feature representations for image Super-Resolution (SR). Second, a series of RCBs are cascaded in a coarse-to-fine manner to construct the Pyramid Residual Clique Network (PRCN), which simultaneously reconstructs multiple SR results (e.g. 2×, 4×, and 8×) in one feed-forward pass. Third, we utilize the human parsing map as the shape prior, and the high-frequency sub-bands of Uniform Discrete Curvelet Transform (UDCT) as the texture prior to enhance the details of reconstructed human body image. Experimental results demonstrate that our proposed method achieves state-of-the-art performance with superior visual quality and PSNR/SSIM scores. Moreover, we show that our results can considerably enhance the performance of human parsing and pose estimation tasks.

Keywords: Human body super-resolution; Residual clique block; Pyramid residual clique network; Uniform discrete curvelet transform

1. Introduction

People are ubiquitous in photographs and surveillance systems, which makes person-related vision techniques, such as pose estimation [1], human parsing [2, 3], and person re-identification [4] important for automatic understanding of media contents. However, the performance of these techniques drops dramatically when the input resolution is very low. To overcome this problem, a typical strategy is to up-sample Low Resolution (LR) images using standard interpolation methods, which inevitably introduce blurriness. Alternatively, image Super-Resolution (SR) are introduced to reconstruct High Resolution (HR) images from LR images with high visual quality.

Recently, Convolutional Neural Network (CNN)-based methods have demonstrated superior performance in image SR. Typically, a CNN based SR method is composed of two crucial components: the feature extraction module and up-sampling module, each of which has a high impact on the SR performance. The feature extraction module usually contains a set of identical feature extraction blocks. Some previous works [5, 6] show that deeper CNNs lead to better performance; however, they consume too much memory and computational time. In this paper, we aim to reach better performance at a lower computational cost. We propose an efficient feature extraction block, namely Residual Clique Block (RCB), which combines clique block with residual connection to allow rich information flow between lower and higher layers and to reduce parameters by making the layers recursive in one clique. The up-sampling module is used to increase resolution and reconstruct the SR image. Inspired by [7, 8, 9], we adopt the deconvolutional layer to build the up-sampling module and construct an efficient pyramid network based on a set of cascaded RCBs, named Pyramid Residual Clique Network (PRCN). The advantage of PRCN is to reconstruct the HR image in a coarse-to-fine manner so as to ease the learning procedure and enhance the performance. As a side product, our model is able to generate multiple SR results (2×, 4×, and 8×) simultaneously in one feed-forward pass.

When compared to general image SR for arbitrary objects, human body image SR is canonical and thus many prior knowledge can be exploited. First, the human parsing maps (i.e. segmentation masks) for different body components, e.g. hair and arm, which can provide an explicit description of the shape of human body. This knowledge is helpful to handle the edges for the task of human body SR. As the human body components are relatively fixed and limited, they can be estimated more easily than arbitrary objects. Second, capturing the details of texture characterized by finer edges is challenging in the spatial domain, but easier in the frequency domain. For example, the wavelet transform (WT) [10, 11] and Non-subsampled Shearlet Transform (NSST) [12, 13] have been successfully applied to represent texture details through their high frequency subbands. In comparison to WT and NSST, the Uniform Discrete Curvelet Transform (UDCT) [14, 15] has more compact frequency response and lower computational cost, making it more

Email address: liuyunan@dlmu.edu.cn (Yunan Liu)

practical for use in applications. Therefore, in this work, we employ human parsing and UDCT to extract shape and texture priors for human body SR, respectively.

In summary, our work makes the following main contributions: (1) We propose a new feature extraction block, namely RCB, which fully utilizes the hierarchical features from LR images. Using RCB, we construct an efficient PRCN that estimates human body prior and reconstructs HR images in a coarse-to-fine manner. (2) Our model estimates two types of human body priors: human parsing maps as a shape prior and high-frequency sub-bands of UDCT as a texture prior. These priors are then used to improve SR performance (3) Experimental results demonstrate the superiority of our method over state-of-the-art methods in terms of PSNR and SSIM. Moreover, we verify the quality of SR images through pose estimation and human parsing tasks, where we observe consistent improvements.

Rest of the paper is organized as follows: Section 2 briefly reviews the related works on the CNN based and prior knowledge based methods, respectively. Section 3 describes our proposed method in detail. Ablation study and comparisons with the state-of-the-art methods are presented in Section 4. Finally, concluding remarks are made in Section 5.

2. Related Work

In this paper, we propose a novel CNN-based method that utilizes prior knowledge to address the problem of super-resolving human body images. Consequently, we will review related works on both CNN-based and prior knowledge-based SR methods.

2.1. CNN Based SR Methods

Since Dong et al. [16] pioneered the first CNN based image SR method (SRCNN), numerous efforts have been made to enhance the learning capability of image SR. Typically, a CNN-based SR methods includes two important components: the feature extraction module and up-sampling module.

For the feature extraction module, early works [16, 17, 18] utilized a straightforward CNN architecture with plain connections to learn the LR-HR mapping. Inspired by the success of residual learning [19] in image classification, some researchers [5, 20, 21, 22] utilized residual connections to tackle the challenge of training deep SR models. Furthermore, some studies employed dense skip connections as the foundation to construct basic feature extraction blocks, such as dense block [23, 24] and residual dense block [6, 25, 26], leading to a significant breakthrough in reconstruction performance. These methods strive to obtain more compact and concise feature representations by constructing deeper feature extraction modules. From the first CNN-based method SRCNN (which comprises only 3 convolutional layers) to the more advanced RCAN [22] (which has over 400 layers), the overall performance has improved dramatically. However, the high computational requirements pose a challenge for deploying deep models on resource-constrained devices in real-world scenarios.

For the up-sampling module, early works [16, 27, 28] utilized bicubic interpolation to pre-upsample the LR images before feeding them into the deep networks, resulting in a significant increase in computational costs. Alternatively, some methods conducted up-sampling after the feature extraction module. To alleviate the burden on the network, some methods [29, 5] utilized sub-pixel convolutional layers, while others employed deconvolutional layers [18, 25, 12]. To reduce the number of model parameters, some methods [30, 31, 11] made use of the recursive strategy for weight sharing across layers at different depths. Nonetheless, these methods lack inflexibility as they require developing specific models tailored to different SR scale factors.

Drawing inspiration from [7, 8], we propose a new PRCN that has the capability to generate SR images with multiple scale factors (i.e. 2×, 4×, and 8×) in a single feed-forward pass. In the feature extraction module, we stack a set of RCB to extract features from different scales. We will describe the detail of PCRN and RCB in sections 3.1 and 3.2, respectively.

2.2. Prior Knowledge based SR Methods

Face image SR is a unique problem within the field of image SR, where facial priors are frequently utilized to improve the quality of reconstructed images. For example, Zhu et al. [32] proposed a deep bi-network specifically for super-resolving LR face images. Song et al. [33] enhanced the reconstruction of facial components by transferring structural details. However, these approaches separate prior prediction and face SR optimization, which makes them difficult to both train and apply.

Alternatively, some other methods leveraged prior knowledge in an end-to-end manner. Bulat et al. [34] presented a unified framework that addresses face SR and alignment simultaneously. Chen et al. [35] developed an end-to-end trainable network that employs facial landmark heatmaps and parsing maps to reconstruct HR face images. Yu et al. [36] adopted a multitask learning strategy for face image SR, which incorporates the facial structural information explicitly into the SR processes. Subsequently, Yu et al. [37] proposed a framework for hallucinating LR face images, where the reconstructed HR face images can be adjusted by tuning specific facial attributes. Grm et al. [38] proposed a face hallucination method that incorporates identity priors into the learning procedure. Zhang et al. [39] introduced a new framework that integrates facial boundaries into nonlinear LR-to-HR mapping at different stages. Lu et al. [40] proposed a split-attention network that enhanced the consistency of facial structure and the fidelity of facial details. Wang et al. [41] proposed a Face-Former method to maintain the facial structure consistency while restoring local facial details. Lu et al. [42] proposed a novel pre-prior guided method that extracts facial prior knowledge from the HR space, enabling the network to obtain sufficient prior knowledge before reconstruction.

Similar to face image, the human body image also exhibits unique variations in shape and texture. Liu et al. [13] utilized the sub-bands of NSST to represent texture details of human body. Compared with NSST, the UDCT has many advantages, such as compact frequency response and lower computational

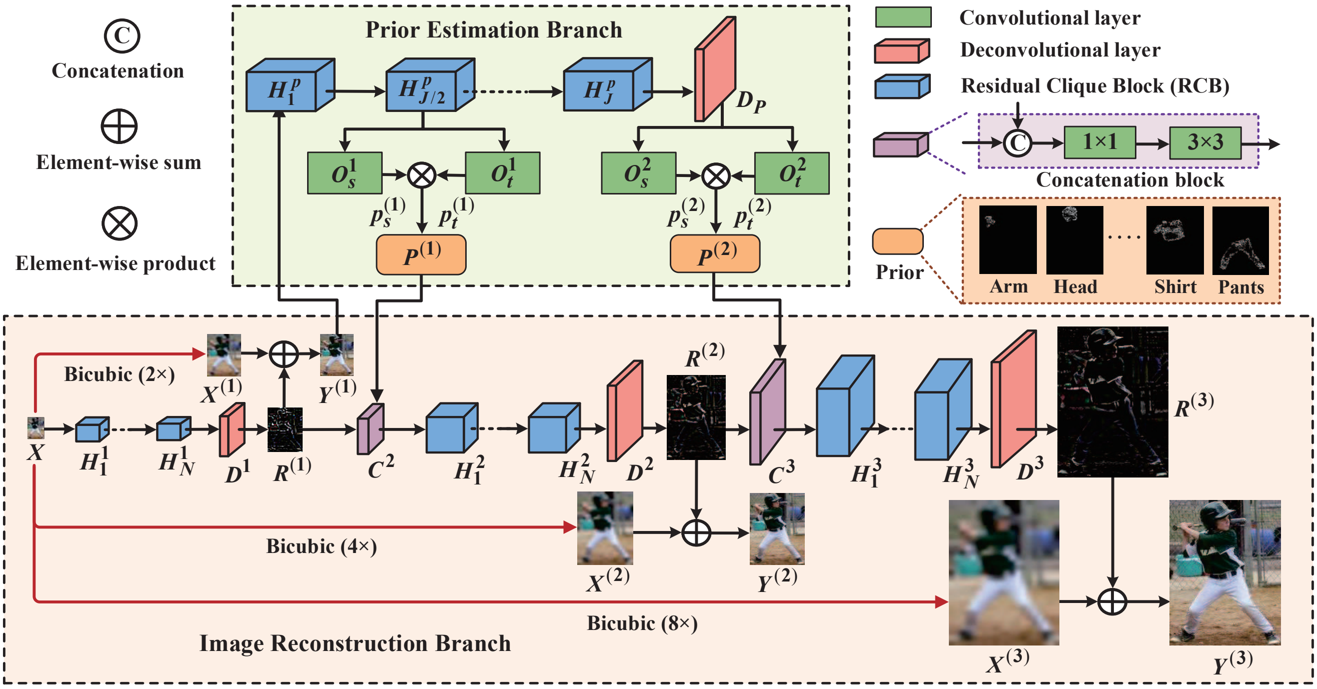


Figure 1: Overview of the proposed method. Our proposed framework consists of prior estimation branch and image reconstruction branch, which progressively reconstruct multiple SR results simultaneously in one feed-forward pass. More details are described in Section 3.1. (In this paper, all figures are best viewed in color.)

complexity, making it a more suitable choice for practical applications. Therefore, in this paper, we use UDCT to represent the texture prior of human body. We will discuss the details of the proposed human body prior in Section 3.3.

3. Proposed Method

In this section, we will first provide an overview of our network architecture. Then, we will introduce our basic feature extraction block in detail, followed by a description of the human body prior.

3.1. Network Architecture

In this paper, we propose a human body SR method based on CNN and prior knowledge. Specifically, our method consists of image reconstruction and prior estimation branches, as illustrated in Fig. 1. The image reconstruction branch takes a LR human body image as input and progressively reconstructs SR results at different pyramid levels. The human body prior estimation branch estimates the human body priors and injects them back into the image reconstruction branch as useful cues. Our model can be trained in an end-to-end manner, and the above two branches collaborate closely with each other. The loss function $\mathcal{L}$ of our framework is defined as follows:

$$\mathcal{L}(\hat{\theta}) = \alpha \mathcal{L}_{Image}(\theta_1) + \mathcal{L}_{Prior}(\theta_2), \tag{1}$$

where α is a hyper-parameter, θ indicates the model parameters to be optimized, $\mathcal{L}_{Image}$ and $\mathcal{L}_{Prior}$ are the loss functions of image reconstruction and prior estimation, respectively.

Image Reconstruction Branch. The image reconstruction branch is a S-level PRCN structure (i.e. S=3). Each pyramid level of PRCN consists of a feature extraction module and a up-sampling module. The feature extraction module consists of a set of cascaded RCBs, which is used for exploiting hierarchical features from LR images. The up-sampling module containing a transposed layer is used for upsampling the extracted feature maps and predicting the residual images. In the image reconstruction branch, we progressively predict results of higher pyramid levels from lower ones, so as to increase the non-linearity of the whole network.

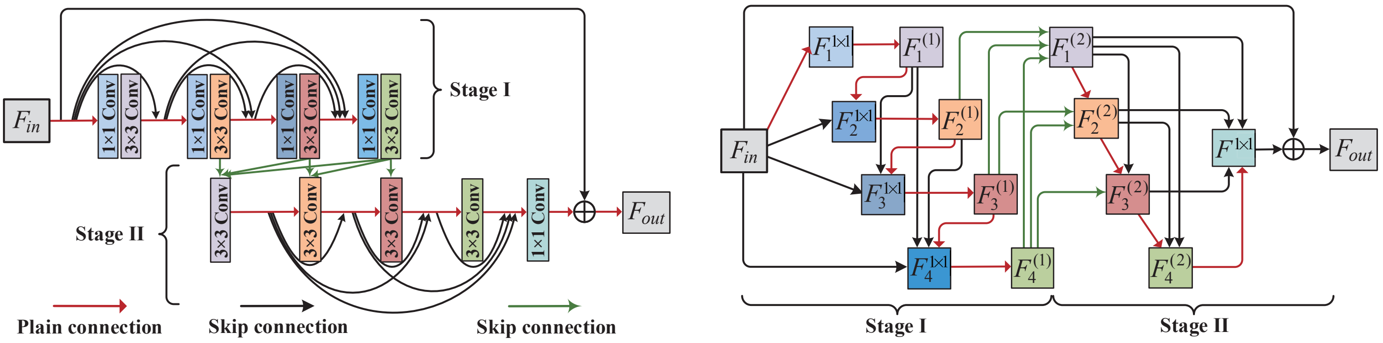
Let X denote the input image, $R^{(s)}$ and $Y^{(s)}$ are predicted residual and reconstructed images at the s-th level, respectively. Suppose we have N RCBs at the s-th level, the output of the n-th RCB can be written as follows:

$$F_n^s = H_n^s(F_{n-1}^s) = H_n^s(H_{n-1}^s(\cdots(H_1^s(X))\cdots)), \tag{2}$$

where H_n^s (n=1, 2, ..., N) indicates the operation function of the n-th RCB at the s-th level. At the first pyramid level, we use the bicubic interpolation operation to increase the resolution of X by a scale factor of T, thus obtaining the upsampled image $X^{(1)}$. Then, we perform element-wise summation to combine $X^{(1)}$ and $R^{(1)}$ as follows,

$$Y^{(1)} = X^{(1)} + R^{(1)} = X^{(1)} + D^1(F_N^1), \tag{3}$$

where D^1 denotes the transposed layer at the first level, which is used to increase the size of feature maps by a factor of 2. It is noted that the upsampling scale factor T of $X^{(s)}$ and the pyramid level s have the relationship: T=2^s. At the following levels,



(a)

(b)

Figure 2: Illustrations of our residual clique block (RCB). (a) The structure of RCB, where F_{in} and F_{out} indicate the input and output of RCB, and the 3×3 convolutional layers with similar colors at different stages share model parameters. In each RCB, we allow information to be passed from low-level layer to high-level layer using the red arrows and black arrows, and transfer information from high-level to low-level layer using the green arrows. (b) The information flow of RCB, where the arrow is the convolution operation, and F indicates the feature maps extracted from convolutional layers.

we first use a concatenation block to extract features from human prior and the residual image at the previous pyramid level. Then, we can obtain reconstructed images $Y^{(2)}$ and $Y^{(3)}$ at the second and third level as follows,

$$Y^{(2)} = X^{(2)}+D^{2}(H_N^2(\cdots(H_1^2(C^2(P^{(1)},R^{(1)})))\cdots)), \quad (4)$$

$$Y^{(3)} = X^{(3)}+D^{3}(H_N^3(\cdots(H_1^3(C^3(P^{(2)},R^{(2)})))\cdots)), \quad (5)$$

where $P^{(1)}$ and $P^{(2)}$ are the prior knowledge estimated from the prior estimation branch, C^2 and C^3 denote the operation of concatenation block at the second and third level, respectively.

In the image reconstruction branch, we supervise the intermediate output at each pyramid level and minimize the combination of loss functions $\mathcal{L}_{Image}$ as follows:

$$\mathcal{L}_{Image}(\theta_1) = \sum_{s=1}^{S=3}\sum_{k=1}^{K} \|Y_k^{(s)} - \hat{Y}_k^{(s)}\|_1, \quad (6)$$

where K is the number of training samples, $Y^{(s)}$ and $\hat{Y}^{(s)}$ denote the reconstruction image and ground-truth HR image at the s-th pyramid level, respectively.

Human Prior Estimation Branch. The main goal of this branch is to predict shape and texture priors simultaneously. These priors are incorporated into the image reconstruction branch to boost performance. Since the distribution of shape and texture features is similar, all features are shared between the two tasks, except for the final two layers at each pyramid level. Specifically, we use two 1×1 convolution layers to predict two kinds of prior knowledge, respectively. Suppose we have J RCBs in the prior estimation branch, the predicted texture prior $P_t^{(s)}$ and shape prior $P_s^{(s)}$ can be written as follows,

$$P_t^{(1)} = O_t^1(H_{J/2}^p(\cdots(H_1^p(Y^{(1)}))\cdots)), \quad (7)$$

$$P_s^{(1)} = O_s^1(H_{J/2}^p(\cdots(H_1^p(Y^{(1)}))\cdots)), \quad (8)$$

$$P_t^{(2)} = O_t^2(D_P(H_J^p(\cdots(H_1^p(Y^{(1)}))\cdots))), \quad (9)$$

$$P_s^{(2)} = O_s^2(D_P(H_J^p(\cdots(H_1^p(Y^{(1)}))\cdots))), \quad (10)$$

where H_j^p ($j = 1, 2, ..., J$) indicates the operation function of the j-th feature extraction block in prior estimation branch, O_t and O_s denote the operation of 1×1 convolutional layer for texture and shape prior estimation, respectively. Following pervious work [13], we combine $P_t^{(s)}$ and $P_s^{(s)}$ by element-wise multiplication as follows,

$$\{P^{(1)}\}_{m=1}^{M} = P_t^{(1)} \otimes \{P_s^{(1)}\}_{m=1}^{M}, \quad (11)$$

$$\{P^{(2)}\}_{m=1}^{M} = P_t^{(2)} \otimes \{P_s^{(2)}\}_{m=1}^{M}, \quad (12)$$

where M denotes the number of channels (i.e., the number of semantic categories of P_s). Our prior estimation branch is able to produce both texture and shape prior knowledge with scale factors 2× and 4×, simultaneously. As presented in Eqs. 4 and 5, the estimated priors $P^{(1)}$ and $P^{(2)}$ will be injected into the image reconstruction branch. The loss functions $\mathcal{L}_{Prior}$ of prior estimation branch can be formulated as:

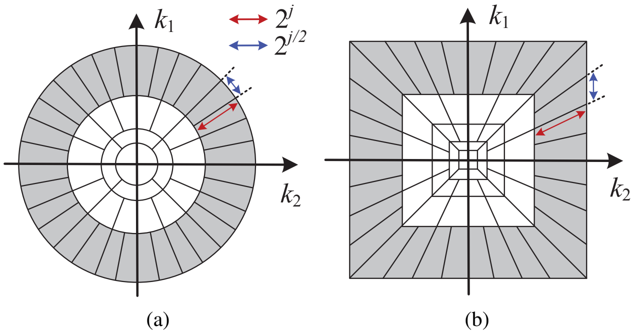
$$\mathcal{L}_{Prior}(\theta_2) = \sum_{s=1}^{S=2}\sum_{k=1}^{K} \|P_k^{(s)} - \hat{P}_k^{(s)}\|_1, \quad (13)$$

where $P^{(s)}$ and $\hat{P}^{(s)}$ denote the estimated prior map and ground-truth label at the s-th pyramid level, respectively.

3.2. Residual Clique Block (RCB)

In recent years, different blocks have been proposed for feature extraction and achieved significant improvement on performance. Zhang et al. [6] introduced residual dense block (RDB) as the basic block for image SR. Yang et al. [43] recently proposed Clique net for image classification task. Different from previous network structures, the convolution layers in the same clique block are updated in a loop manner, so that the information flow among layers is maximized.

Inspired by the previous works [6, 43], we propose a new feature extraction block for image SR task called residual clique block (RCB). Similar to the Clique block, the propagation of



(a)

(b)

Figure 3: Support of $\varphi_{j,l,\boldsymbol{k}}$ in the frequency domain. (a) The tiling of continuous UDCT. (b) The tiling of discrete UDCT.

our RCB contains two stages, and the convolution layers in each block are constructed as a clique. In contrast to the Clique block, we adopt 1×1 convolution layer before each 3×3 convolution layer at the first stage so that the number of input feature maps is the same as the corresponding 3×3 convolution layer in the second stage. This design allows the weights of the corresponding 3×3 convolutional layers (blocks with different colors and labels in Fig. 2a) at different stages to be shared, updating the convolution layers in our RCB in a novel loop manner. We also incorporate the local feature fusion and local residual learning into our RCB structure. Local feature fusion refers to a 1×1 convolution layer that adaptively fuses refined information from the convolutional layers in the second stage of RCB. Then, we apply local residual learning to further improve the network's representation ability.

Fig. 2b illustrates the information flow of RCB, which involves an alternate optimization of feature maps in a loop. At the second stage, the output of one feature map serves as the input of another, facilitating the maximization of information flow. Let F_{in} and F_{out} denote the input and output of RCB, $F_i^{1\times1}$ $(i = 1, ..., 4)$ denotes the feature maps extracted from i-th 1×1 convolutional layer, $F_i^{(1)}$ and $F_i^{(2)}$ denote the feature maps extracted from i-th 3×3 convolutional layer at the first and second stage, respectively. At the first stage, the feed-forward pass of feature maps can be formulated as follows,

$$F_i^{(1)} = \sigma(\mathcal{W}_i^{(i)} * F_i^{1\times1}), \tag{14}$$

$$F_i^{1\times1} = \sigma(\sum_{j=1}^{i-1} \mathcal{W}_i^{1\times1} * F_j^{(1)} + \mathcal{W}_i^{1\times1} * F_{in}), \tag{15}$$

where $*$ is convolution operation, $\mathcal{W}$ indicates the weights of corresponding convolutional layer, σ stands for activation function. At the second stage, the feed-forward pass of feature maps can be written as follows,

$$F_i^{(2)} = \sigma(\sum_{j=1}^{4-i} \mathcal{W}_i^{(2)} * F_{i+j}^{(1)} + \sum_{j=1}^{i-1} \mathcal{W}_i^{(2)} * F_j^{(2)}). \tag{16}$$

Finally, we obtain the output of RCB using the local feature fusion and local residual learning as follows,

$$F_{out} = \sigma(\mathcal{W}^{1\times1} * [F_1^{(2)}, F_2^{(2)}, F_3^{(2)}, F_4^{(2)}]) + F_{in}, \tag{17}$$

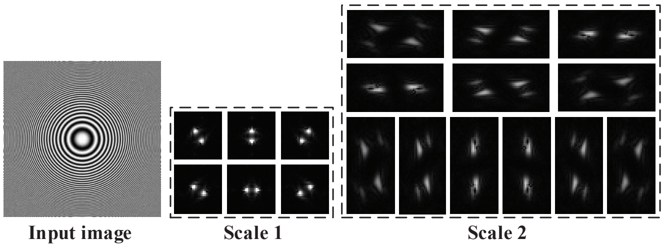


Figure 4: The high-frequency sub-bands of 2-level UDCT, which can capture the texture details of the input image across different scales and directions.

where $[\cdot]$ is the concatenation operation.

3.3. Human Prior Knowledge

In this paper, we aim to enhance the visual quality of reconstructed human body images by integrating two types of priors into the SR process. The first type of prior uses human parsing maps as shape priors, which provides global structural information of the human body. The parsing maps contain different components for different human bodies, such as hat, hair, face, etc., which contribute to the semantic layout information used for reconstruction. The second type of prior employs UDCT, an effective multi-resolution analysis method, to capture human texture priors. UDCT decomposes the images into several high-frequency sub-bands are obtained, each one capturing detailed features of a particular direction of the image. We use element-wise summation to fuse all the high-frequency sub-bands, creating an omnidirectional texture prior that represents the local texture details of human body images. Previous studies have explored the use of wavelet transform (WT) for image SR in deep networks [10, 44]. However, a common limitation of WT is its inability to accurately represent the curves and edges of two-dimensional images due to its isotropic property. To overcome this disadvantage, other multi-resolution analysis methods such as NSCT [45, 46] and NSST [12, 13] have been proposed for image SR. When compared to the WT, NSCT and NSST domains, the UDCT used in the paper offers many advantages, including better frequency response, lower coefficient redundancy, and lower computational cost. These benefits make UDCT more suitable for representing the texture details of human body images.

Suppose $\boldsymbol{t}$=$\{t_1, t_2\}$∈$\mathbb{R}^2$ and $\boldsymbol{\omega}$=$\{\omega_1, \omega_1\}$ are the variables in spatial domain and frequency domain, respectively. The basis function of curvelet transform is indexed by scale j, direction l and position $\boldsymbol{k}$=$\{k_1, k_2\}$, the 2-D curvelet transform φ_μ can be written as follows,

$$\varphi_\mu = \varphi_j(R_{\theta_L}(\boldsymbol{t} - \boldsymbol{t}_{\boldsymbol{k}}^{(j,l)})), \tag{18}$$

where $\boldsymbol{\mu}$=$(j, l, \boldsymbol{k})$ indicates the index set, R_{θ_L} is rotation by angle of θ_L. Fig. 3a shows the frequency support of continuous curvelet transform, where the center of the curvelet is a rotated grid $\boldsymbol{t}_{\boldsymbol{k}}^{(j,l)}$=$R_{\theta_L}^{-1}(k_1 2^j, k_2 2^{j/2})$. Then, the coefficients of

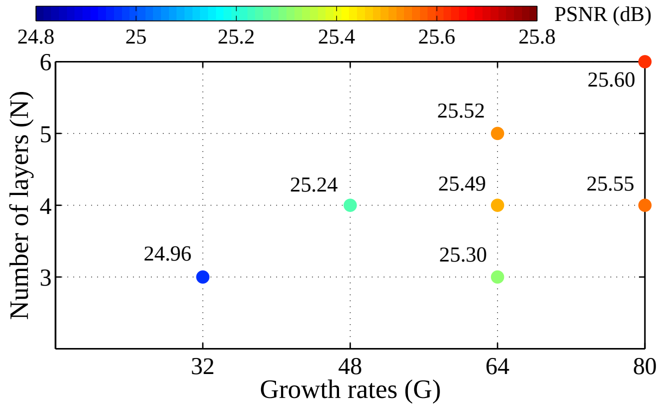


Figure 5: The PSNR results under different G and N on 8× Human body SR. The performance improves as G and N increase.

curvelet transform can be computed by the inner product between $f(t_1, t_2) \in L^2(\mathbb{R}^2)$ and curvelet transform $\varphi_\mu(\boldsymbol{t})$ as follows,

$$c_\mu = \langle f(\boldsymbol{t}), \varphi_\mu(\boldsymbol{t}) \rangle = \int f(\boldsymbol{t}) \varphi_\mu(\boldsymbol{t}) d\boldsymbol{t}. \tag{19}$$

Since the family of smooth windows is used to define the basic function of UDCT, it is important to establish the equivalent operation in the frequency domain for a discrete image. Fig. 3b shows the frequency support of discrete curvelet transform. Let $u_l(\boldsymbol{\omega})$ denote a family of $2N$+1 smooth windows, where l=0, 1, ..., 2N and N is an integer, the set of curvelet function must obey the following rules: (1) All windows functions are 2π periodic in ω_1 and ω_2. (2) The first window $u_0(\boldsymbol{\omega})$ (i.e. lowpass window) has a square-shaped support and zero outside $[-\pi/2, \pi/2]^2$, while other $2N$ windows have wedge-shaped supports. (3) The widest part of the wedge-shaped support of $u_l(\boldsymbol{\omega})$, $l \neq 0$, should smaller than π, and the sum of $u_0^2(\boldsymbol{\omega})$ and $u_l^2(\boldsymbol{\omega}) + u_l^2(-\boldsymbol{\omega})$ equal to 1.

Let us define a 7-band filter bank defined from 7 $u_l(\boldsymbol{\omega})$ windows as follows,

$$F_0(\boldsymbol{\omega}) = 2u_0(\boldsymbol{\omega}), F_l(\boldsymbol{\omega}) = 2^{\frac{n+3}{2}} u_l(\boldsymbol{\omega}). \tag{20}$$

Suppose $\hat{F}_{j,l}(\boldsymbol{\omega})$ denote the equivalent filters at scale j and direction l, we have

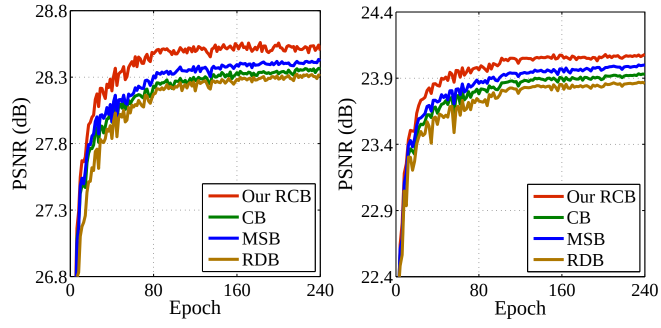
$$\hat{F}_{j,l}(\boldsymbol{\omega}) = F_l^{(j)}(2^{J-j}\boldsymbol{\omega}) \prod_{i=0}^{J-j-1} F_0^{(J-i)}(2^i \boldsymbol{\omega}). \tag{21}$$

Since $F_0^{(i)}(\boldsymbol{\omega})$, i=1, ..., J are defined from $u_0(\boldsymbol{\omega})$ in Eq. 20, the equivalent high-pass filter is formulated as follows,

$$\hat{F}_{j,l}(\boldsymbol{\omega}) = \begin{cases} 2^{J-j-1} F_l^{(j)}(2^{J-j}\boldsymbol{\omega}) F_0^{j+1}(2^{J-j-1}\boldsymbol{\omega}), & |\omega_1|, |\omega_2| < \pi/2^{J-j} \\ 0, & \pi/2^{J-j} \leq |\omega_1|, |\omega_2| < \pi. \end{cases} \tag{22}$$

Also, the equivalent low-pass filter is written as follows,

$$\hat{F}_0(\boldsymbol{\omega}) = \begin{cases} 2^{J-1} F_0(2^{J-1}\boldsymbol{\omega}), & |\omega_1|, |\omega_2| < \pi/2^J \\ 0, & pi/2^J \leq |\omega_1|, |\omega_2| < \pi. \end{cases} \tag{23}$$



(a) (b)

Figure 6: Comparison of different feature extraction blocks, i.e. Residual Dense Block (RDB) [6], Clique Block (CB) [43], Multi-Scale Block (MSB) [13], and our RCB with a scale factor of (a) 4× and (b) 8× on the HumanSR dataset.

Table 1: Comparison of different connection strategies on 8× human body image SR.

Plain connection	Skip connection (Low-to-high)	Skip connection (High-to-low)	PSNR
√			24.95
√	√		25.23
√		√	25.17
√	√	√	25.49

The filter bank of UDCT at j-th level has $2N_j$ directions, where N_j=3×2^j. As shown Fig. 4, the zoneplate image is decomposed by a 2-level UDCT, where we can obtain 6 and 12 high-frequency subbands from finer to coarser scales. We find that the high-frequency sub-bands of UDCT can effectively capture texture details through different high-pass filters. To better represent the global and local details of the human body, we integrate human parsing with UDCT to extract human body knowledge.

4. Experiments

In this section, we first provide the detail of experimental settings. Then, we study the contribution of our RCB and prior knowledge by the ablation experiments. Finally, we compare our method with state-of-the-art methods.

4.1. Settings

Dataset. We conducted experiments to evaluate the performance on the HumanSR dataset [13]. The human body images are collected from three public datasets: the ATR [47], CIHP [48], and LIP [49] datasets. The training set of HumanSR dataset consists of 30,000 images, of which 6,500 are from CIHP, 7500 are from LIP, and the rest come from ATR. The test set consists of 600 images, with 200 images from each of the three datasets, i.e., ATR, CIHP, and LIP.

Implementation details. We resize the human body images to 320×160 as ground truth HR images and generate LR input images by applying bicubic downsampling on HR images.

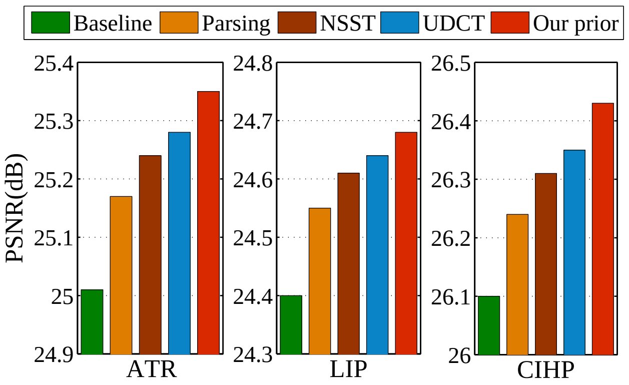


Figure 7: Comparison of different prior knowledge on humanSR test set with a scale factor of 8×.

After that, the training samples are augmented by horizontally flipping them. To generate the ground truth of texture prior, we apply 1-level UDCT on the HR image,which generates 6 high frequency sub-bands. In addition, we adopt the HIPN model [3] to estimate the ground truth shape prior with 19 semantic classes. Each pyramid level in both image reconstruction and prior estimation branches contains 2 RCBs. The α in Eq. 1 is set to 0.5. The deconvolutional layers adopt 17×17 filters for all scaling factors, and each convolution layer in RCB is followed by a Leakly Rectified Linear Unit (LReLU) except for the local fusion layers, in which the negative scope is set to 0.05. The learning rate is initially set to 0.0001 for all layers and is halved every 100 epochs. The total training epochs are set to 240. We use Adam as our optimizer and conduct all experiments using Pytorch on two NVIDIA RTX2080Ti GPUs.

Evaluation metrics. Following the standard protocol, we use the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [50] as the evaluation indicators. Additionally, we utilize feature similarity (FSIM) [51], visual information fidelity (VIF) [52] and learned perceptual image patch similarity (LPIPS) [53] to further evaluate the reconstruction performance of different methods. Regarding human parsing and pose estimation, we follow previous works [49, 13] that adopt the intersection-over-union (IoU) and percentage of correct keypoints with respect to head (PCKh) as the evaluation metrics, respectively. It is worth noting that a lower value of the LPIPS metric indicates a higher perceptual similarity of the reconstructed images, while higher values for other metrics indicate better performance.

4.2. Ablation Study

Evaluation on Key Settings of RCB. We first investigate the influence of the growth rates (i.e., the number of 3×3 filters denoted as G) and the number of convolutional layers per RCB (denoted as N). We construct the image reconstruction branch using different combinations of N and G. In Fig. 5, we evaluate the SR performance w.r.t. PSNR on the HumanSR test set. We find that increasing either N or G brings in consistent performance improvements, indicating deeper is better. Considering the trade-off between accuracy and speed, the combination of N=4 and G=64 will be used in the following experiments. Then, we explore the effect of the connection strategy of RCB on SR performance. Table 1 presents an evaluation of different connection strategies on the HumanSR test set. We obtain consistent improvements by using two kinds of skip connections compared to the baseline model that only uses plain connections. When the skip connection for low-to-high passing is removed, we use the plain connection to combine the first stage with the second stage directly. By using three connection strategies simultaneously, we achieve a 0.54 with respect to PSNR. This gain is mainly due to passing richer information through different layers, resulting in more powerful features.

Table 2: Trade-off between performance vs. number of model parameters and running time for handling 2×, 4×, and 8× SR on HumanSR test sets.

Method	Parameters (M)	Time (s)	PSNR (dB)
LMSN + NSST [13]	6.4	0.54	29.64
LMSN + our prior	6.4	0.54	29.73
DBPN [25]	31.5	1.32	29.45
DBPN + our prior	37.7	1.66	29.77
RCAN [22]	48.6	1.62	29.55
RCAN + our prior	54.9	1.95	29.87
Our PRCN (w/o prior)	4.1	0.35	29.51
Our PRCN	6.2	0.47	29.80

Effectiveness of RCB. To validate the effectiveness of our RCB, we conduct comparative with Dense Block (DB) [23], Residual Dense Block (RDB) [6], Clique Block (CB) [43], Multi-Scale Block (MSB) [13] and our RCB on human body image SR. For quick verification, we construct five simple network architectures, each consisting of a feature extraction module and an up-sampling reconstruction module. The up-sampling modules of the five networks are identical to each other, while the feature extraction modules in the five networks contain different feature extraction blocks, and each network contains only one block. All the models are trained on the HumanSR dataset with a scaling factor 8× in the same environment. The PSNR results in Fig. 6 show that our RCB outperforms other feature blocks by a large margin, demonstrating the effectiveness of our RCB.

Impact of Prior Knowledge. We conduct two groups of experiments to validate the effectiveness of our proposed human body prior: (1) In Fig. 7, we first compare the effectiveness of different priors for 4× SR of human body images. To achieve this, we use the prior estimation branch to generate four types of prior knowledge: human parsing maps, UDCT texture prior, NSST texture prior, and our proposed integrated prior. Among these priors, our proposed integrated prior, which combines parsing with UDCT, provides more useful cues for better SR results. (2) We use three state-of-the-art SR methods, i.e., RCAN [22], DBPN [25] and LMSN [13] to construct the image reconstruction branch and integrate them with the predictions of our prior estimation branch. Specifically, we resize the estimated prior maps to match the size of the feature maps of different models. In Table 2, we provide a comprehensive

Table 3: Comparison with state-of-the-art methods on HumanSR test sets. **Bold**/underline indicate the best/second-best results.

Method	Scale	ATR					LIP					CIHP				
		PSNR	SSIM	LPIPS	FSIM	VIF	PSNR	SSIM	LPIPS	FSIM	VIF	PSNR	SSIM	LPIPS	FSIM	VIF
Bicubic	×2	30.82	0.923	0.275	0.938	0.634	30.89	0.919	0.292	0.923	0.616	32.22	0.932	0.266	0.943	0.653
MSLapSRN [8]		34.14	0.942	0.135	0.958	0.689	33.36	0.931	0.146	0.947	0.675	35.35	0.946	0.141	0.960	0.697
RCAN [22]		34.31	0.948	0.128	0.962	0.696	33.54	0.940	0.138	0.955	0.682	35.53	0.954	0.133	0.965	0.707
NLSN [20]		34.37	0.953	0.116	0.969	0.704	33.64	0.946	0.126	0.963	0.695	35.55	0.958	0.128	0.969	0.713
LMSN [13]		34.35	0.951	0.122	0.965	0.699	33.60	0.944	0.131	0.958	0.687	35.58	0.959	0.118	0.971	0.715
Our method		**34.46**	**0.957**	**0.112**	**0.975**	**0.712**	**33.80**	**0.951**	**0.118**	**0.968**	**0.698**	**35.74**	**0.964**	**0.107**	**0.982**	**0.723**
Bicubic	×4	26.05	0.794	0.369	0.828	0.485	25.61	0.758	0.388	0.819	0.469	26.90	0.807	0.351	0.834	0.503
OISR [29]		28.68	0.866	0.234	0.901	0.631	27.86	0.823	0.252	0.892	0.611	29.77	0.884	0.228	0.912	0.652
EDSR [5]		28.80	0.873	0.225	0.907	0.639	27.99	0.825	0.247	0.896	0.615	29.87	0.888	0.224	0.917	0.658
DBPN [25]		28.85	0.871	0.228	0.912	0.644	28.04	0.828	0.242	0.899	0.623	29.90	0.890	0.223	0.920	0.664
MSLapSRN [8]		28.82	0.873	0.225	0.910	0.648	28.05	0.830	0.239	0.903	0.627	29.93	0.890	0.217	0.922	0.662
RCAN [22]		28.92	0.878	0.207	0.919	0.656	28.11	0.832	0.233	0.909	0.631	29.96	0.893	0.205	0.925	0.669
NLSN [20]		28.85	0.875	0.210	0.915	0.650	28.10	0.830	0.236	0.906	0.632	30.01	0.895	0.202	0.928	0.673
LMSN [13]		29.08	0.885	0.203	0.924	0.658	28.23	0.839	0.225	0.914	0.640	30.05	0.897	0.195	0.928	0.676
Our method		**29.20**	**0.891**	**0.194**	**0.931**	**0.667**	**28.37**	**0.844**	**0.213**	**0.921**	**0.648**	**30.21**	**0.902**	**0.186**	**0.937**	**0.683**
Bicubic	×8	22.57	0.657	0.495	0.715	0.284	22.07	0.597	0.503	0.672	0.267	23.21	0.673	0.482	0.737	0.305
OISR [29]		24.74	0.737	0.341	0.794	0.462	24.17	0.683	0.352	0.759	0.440	25.84	0.757	0.333	0.815	0.486
EDSR [5]		24.88	0.742	0.333	0.798	0.469	24.27	0.689	0.344	0.766	0.445	25.91	0.759	0.325	0.820	0.493
DBPN [25]		24.95	0.744	0.329	0.803	0.475	24.35	0.692	0.338	0.770	0.453	25.95	0.761	0.320	0.823	0.496
MSLapSRN [8]		25.01	0.748	0.321	0.808	0.480	24.40	0.694	0.335	0.774	0.456	25.92	0.763	0.315	0.829	0.505
RCAN [22]		25.08	0.750	0.315	0.812	0.483	24.43	0.695	0.329	0.778	0.463	26.06	0.765	0.308	0.836	0.514
NLSN [20]		24.85	0.740	0.336	0.795	0.470	24.31	0.690	0.340	0.763	0.448	25.91	0.757	0.329	0.817	0.490
LMSN [13]		25.20	0.754	0.307	0.819	0.487	24.52	0.699	0.314	0.786	0.474	26.18	0.770	0.296	0.844	0.527
Our method		**25.35**	**0.759**	**0.294**	**0.827**	**0.496**	**24.68**	**0.704**	**0.302**	**0.794**	**0.480**	**26.43**	**0.776**	**0.288**	**0.853**	**0.534**

comparison with other methods on the HumanSR test sets. It is worth noting that the comparative methods [22, 25, 13] requires training three different models for handling 2×, 4× and 8× S-R, respectively, while our PCRN is capable of handling multiple upsampling scales through a single PCRN model. From Table 2, we observe that all methods show significant improvement by using our prior knowledge, and our PCRN outperforms other methods with fewer parameters and faster speed.

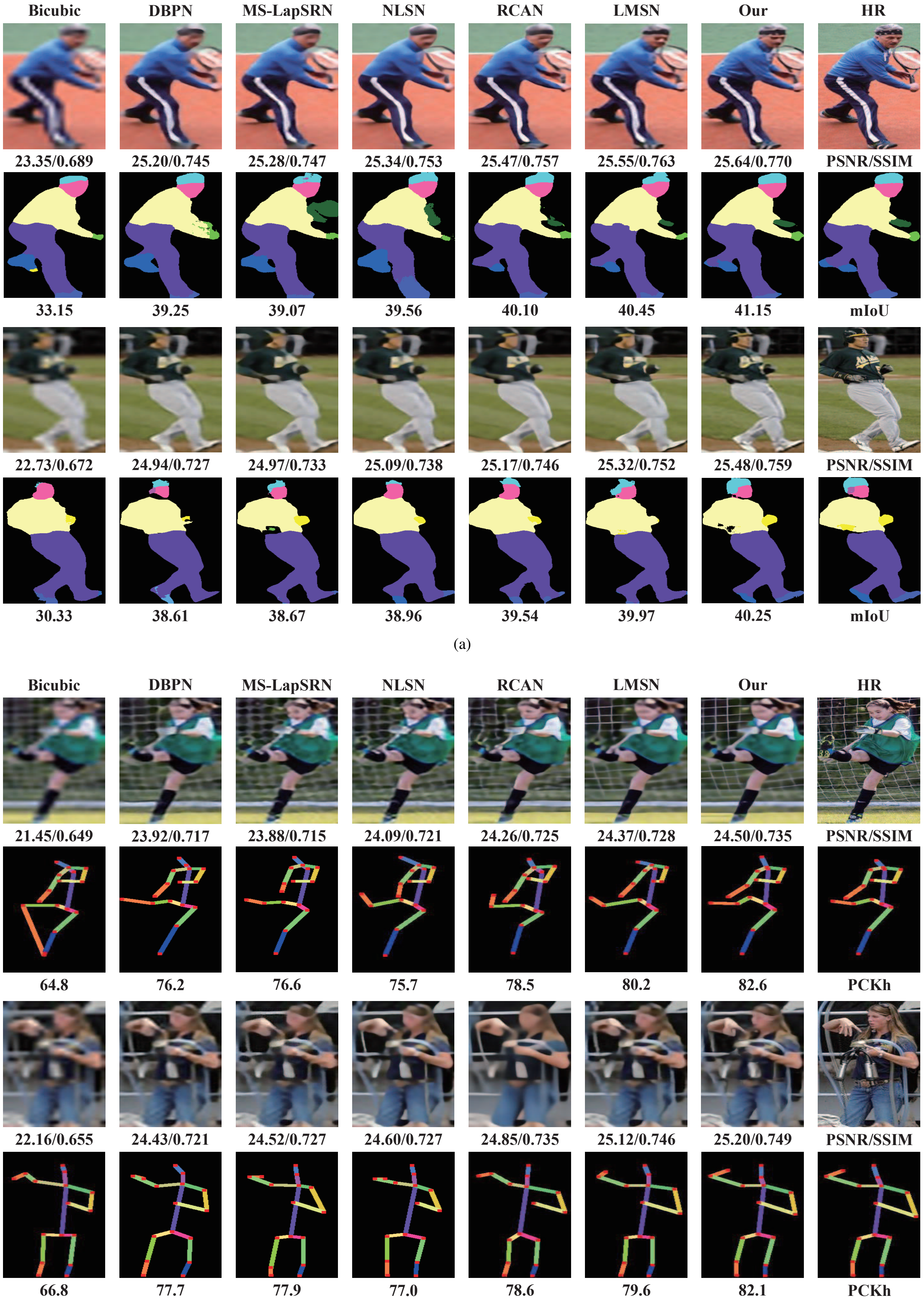
4.3. Comparison with State-of-the-art Methods

In this section, we compare our method with state-of-the-art image SR methods, i.e., EDSR [5], RCAN [22], DBPN [25], MSLapSRN [8], OISR [29], NLSN [20], and LMSN [13]. For fair comparison, we use the released codes of the above methods and re-train all the models on the same HumanSR training set. The PSNR, SSIM, LPIPS, FSIM, and VIF of different methods are reported in Table 3. Apart from traditional SR metrics, we also evaluate the human parsing and pose estimation results on top of reconstructed SR images. For human parsing and pose estimation, we apply the HRNet [54] on top of SR images from HumanSR test sets. The mIoU and PCKh results of different SR methods are also reported in Table 4. As presented in Tables 3 and 4, our method outperforms other methods in terms of SR reconstruction across various test sets and scale factors. Furthermore, we find that the reconstructed images generated by our method lead to superior results in both human parsing and pose estimation compared to other SR methods.

In Fig. 8, we provide the qualitative results of SR, human parsing and pose estimation. As shown in the figure, our

Table 4: Comparison with state-of-the-art methods for human parsing and pose estimation. **Bold**/underline indicate the best/second best results.

Method	Scale	ATR		LIP		CIHP	
		mIoU	PCKh	mIoU	PCKh	mIoU	PCKh
Bicubic	×2	47.63	89.7	59.4	91.9	54.91	91.0
MSLapSRN [8]		49.28	91.0	61.5	93.4	57.30	92.9
RCAN [22]		49.32	91.3	62.0	93.8	57.36	93.2
NLSN [20]		49.57	91.6	62.6	94.5	57.51	93.7
LMSN [13]		49.55	91.6	62.0	94.2	57.59	94.1
Our method		**49.71**	**92.2**	**62.9**	**94.9**	**57.72**	**94.5**
Bicubic	×4	40.15	82.2	54.25	86.7	50.55	85.5
OISR-LF [29]		44.66	85.6	57.50	89.0	54.82	89.0
EDSR [5]		45.25	86.5	57.83	89.3	55.41	89.3
DBPN [25]		45.19	86.0	57.75	89.7	55.33	89.5
MSLapSRN [8]		45.34	86.3	57.85	89.5	55.55	89.7
RCAN [22]		45.52	86.6	57.93	90.1	55.75	90.1
NLSN [20]		45.33	86.0	57.90	90.0	55.88	90.3
LMSN [13]		45.92	87.0	58.12	90.5	55.92	90.5
Our method		**46.02**	**87.4**	**58.60**	**90.9**	**56.03**	**90.9**
Bicubic	×8	27.11	66.5	43.64	71.3	40.60	68.3
OISR [29]		35.65	79.0	53.03	84.6	48.20	81.2
EDSR [5]		36.78	80.2	53.69	85.2	49.01	83.3
DBPN [25]		36.86	79.6	53.02	84.6	48.93	82.5
MSLapSRN [8]		37.32	80.2	53.25	85.2	49.09	82.9
RCAN [22]		37.73	80.9	53.51	86.1	49.42	83.5
NLSN [20]		36.65	79.9	53.05	84.8	48.88	82.4
LMSN [13]		38.25	81.4	53.94	85.8	49.82	84.3
Our method		**38.40**	**81.7**	**54.05**	**86.3**	**49.92**	**84.6**



(a)

(b)

Figure 8: Qualitative results of human body image SR, (a) human parsing, and (b) pose estimation.

method outperforms other methods such as DBPN [25], M-SLapSRN [8], NLSN [20], RCAN [22], and LMSN [13] in handling both overall shape and local texture. In general, our method is able to reconstruct more accurate HR images, which not only exhibit higher visual quality, but also enable more precise analysis of the human body.

5. Conclusions

In this paper, we focus on the task of super-resolving human body images. A coherent framework is proposed to jointly solve the image reconstruction and human prior estimation, to leverage prior knowledge for high-resolution human body reconstruction. Within this framework, we introduce an efficient feature extraction block, namely RCB, which outperforms previous blocks for image SR. Furthermore, we construct a progressive network PRCN by cascading a series of RCBs. Experimental results demonstrate that our method outperforms state-of-the-art methods for human body super-resolution. We also show that by applying our method to LR images, better performance can be achieved for human parsing and pose estimation tasks than other SR methods. Encouraging by this work, future research will focus on designing task-driven SR methods and better integrating image SR methods into other vision tasks to address performance degradation due to the decreased resolution.

Acknowledgment

This work was supported in part by the China Postdoctoral Science Foundation under Grant No. 2022M720624; in part by the Basic Scientific Research Project of Liaoning Provincial Department of Education under Grant No. LJKQZ2021152; in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61602226, 61802212, and 61976124; in part by the Key Research Project on Philosophy and Social Science by the Ministry of Education under Grant No. 21JZD049; in part by the Shandong Provincial Natural Science Foundation No. ZR2019BF017.

References

[1] C. Wang, F. Zhang, X. Zhu, S. Ge, Low-resolution human pose estimation, Pattern Recognition 126 (2022) 108579.

[2] Y. Liu, L. Zhao, S. Zhang, J. Yang, Hybrid resolution network using edge guided region mutual information loss for human parsing, in: ACM International Conference on Multimedia, 2020, pp. 1670–1678.

[3] Y. Liu, S. Zhang, J. Yang, P. Yuen, Hierarchical information passing based noise-tolerant hybrid learning for semi-supervised human parsing, in: AAAI Conference on Artificial Intelligence, 2021, pp. 2207–2215.

[4] G. Wu, X. Zhu, S. Gong, Learning hybrid ranking representation for person re-identification, Pattern Recognition 121 (2022) 108239.

[5] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1132–1140.

[6] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[7] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep laplacian pyramid networks for fast and accurate super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2017, pp. 5835–5843.

[8] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Fast and accurate image super-resolution with deep laplacian pyramid networks, IEEE Transactions on Pattern Analysis Machine Intelligence 41 (11) (2019) 2599–2613.

[9] Y. Tang, W. Gong, X. Chen, W. Li, Deep inception-residual laplacian pyramid networks for accurate single-image super-resolution, IEEE Transactions on Neural Networks and Learning Systems 31 (5) (2020) 1514–1528.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep wavelet prediction for image super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1100–1109.

[11] Z. Li, Z. Kuang, Z. Zhu, H. Wang, X. Shao, Wavelet-based texture reformation network for image super-resolution, IEEE Transactions on Neural Networks and Learning Systerms 31 (2022) 2647–2660.

[12] Y. Liu, S. Zhang, C. Wang, J. Xu, Single image super-resolution via hybrid resolution nsst prediction, Computer Vision and Image Understand 207 (2021) 103202.

[13] Y. Liu, S. Zhang, J. Xu, J. Yang, Y.-W. Tai, An accurate and lightweight method for human body image super-resolution, IEEE Transactions on Image Processing 30 (2021) 2888–2879.

[14] T. T. Nguyen, H. Chauris, Uniform discrete curvelet transform, IEEE Transactions on Signal Processing 58 (7) (2010) 3618–3634.

[15] C. Wang, S. Wang, B. Ma, J. Li, X. Dong, Z. Xia, Transform domain based medical image super-resolution via deep multi-scale network, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019, pp. 2387–2391.

[16] C. Dong, C. C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, 2014, pp. 184–199.

[17] W. Shi, J. Caballero, F. Huszar, J. Totz, A. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.

[18] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European Conference on Computer Vision, 2016, pp. 391–407.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[20] Y. Mei, Y. Fan, Y. Zhou, Image super-resolution with non-local sparse attention, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2021, pp. 3517–3526.

[21] N. Ahn, B. Kang, K. Sohn, Efficient deep neural network for photo-realistic image super-resolution, Pattern Recognition 127 (2022) 108649.

[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: European Conference on Computer Vision, 2018, pp. 294–310.

[23] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: IEEE International Conference on Computer Vision, 2017, pp. 4809–4817.

[24] K. Jiang, Z. Wang, P. Yi, J. Jiang, Hierarchical dense recursive network for image super-resolution, Pattern Recognition 107 (2020) 107475.

[25] M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2018, pp. 1664–1673.

[26] S. Anwar, N. Barnes, Densely residual laplacian super-resolution, IEEE Transactions on Pattern Analysis Machine Intelligence 44 (3) (2022) 1192–1204.

[27] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.

[28] Y. Tai, J. Yang, X. Liu, C. Xu, Memnet: A persistent memory network for image restoration, in: IEEE International Conference on Computer Vision, 2017, pp. 4549–4557.

[29] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, J. Cheng, ODE-inspired network design for single image super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2019, pp. 1732–1741.

[30] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual

network, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2017, pp. 2790–2798.

[31] J. Kim, J. Lee, K. Lee, Deeply-recursive convolutional network for image super-resolution, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2016, pp. 1637–1645.

[32] S. Zhu, S. Liu, C. C. Loy, X. Tang, Deep cascaded bi-network for face hallucination, in: European Conference on Computer Vision, 2016, pp. 614–630.

[33] Y. Song, J. Zhang, S. He, L. Bao, Q. Yang, Learning to hallucinate face images via component generation and enhancement, in: International Joint Conference on Artificial Intelligence, 2017, pp. 4537–4543.

[34] A. Bulat, G. Tzimiropoulos, Super-fan: integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2018, pp. 109–117.

[35] Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, Fsrnet: End-to-end learning face super-resolution with facial priors, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2018, pp. 2492–2501.

[36] X. Yu, B. Fernando, F. Porikli, R. Hartley, Face super-resolution guided by facial component heatmaps, in: European Conference on Computer Vision, 2018, pp. 219–235.

[37] X. Yu, B. Fernando, R. Hartley, F. Porikli, Semantic face hallucination: super-resolving very low-resolution face images with supplementary attributes, IEEE Transactions on Pattern Analysis Machine Intelligence 42 (11) (2020) 2926–2943.

[38] K. Grm, W. J. Scheirer, V. Struc, Face hallucination using cascaded super-resolution and identity priors, IEEE Transactions on Image Processing 29 (2020) 2150–2165.

[39] Y. Zhang, Y. Wu, L. Chen, Msfsr: A multi-stage face super-resolution with accurate facial representation via enhanced facial boundaries, in: IEEE Confefence on Computer Vision and Pattern Recognition Workshops, 2020, pp. 2120–2129.

[40] T. Lu, Y. Wang, Y. Zhang, Y. Wang, W. Liu, Z. Wang, J. Jiang, Face hallucination via split-attention in split-attention network, in: ACM International Conference on Multimedia, 2021, pp. 5501–5509.

[41] Y. Wang, T. Lu, Y. Zhang, Z. Wang, J. Jiang, Z. Xiong, Faceformer: Aggregating global and local representation for face hallucination, IEEE Transactions on Circuits and Systems for Video Technology (Early access) (2022).

[42] T. Lu, Y. Wang, Y. Zhang, J. Jiang, Z. Wang, Z. Xiong, Rethinking prior-guided face super-resolution: A new paradigm with facial component prior, IEEE Transactions on Neural Networks and Learning Systems (Early access) (2022).

[43] Y. Yang, Z. Zhong, T. Shen, Z. Lin, Convolutional neural networks with alternately updated clique, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2018, pp. 2413–2422.

[44] Z. Li, Z. Kuang, Z. Zhu, H. Wang, X. Shao, Wavelet-based texture reformation network for image super-resolution, IEEE Transactions on Image Processing 31 (2022) 2647–2660.

[45] Y. Sang, J. Sun, S. Wang, K. Li, H. Qi, Medical image super-resolution via granular multi-scale network in nsct domain, in: IEEE International Conference on Multimedia and Expo, 2020, pp. 1–6.

[46] Y. Sang, J. Sun, S. Wang, Y. Peng, X. Zhang, Z. Yang, Multi-scale information distillation network for image super resolution in nsct domain, in: International Conference on Neural Information Processing, 2019, pp. 50–59.

[47] X. Liang, S. Liu, X. Shen, et al., Deep human parsing with active template regression, IEEE Transactions on Pattern Analysis Machine Intelligence 37 (12) (2015) 2402–2414.

[48] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, L. Lin, Instance-level human parsing via part grouping network, in: European Conference on Computer Vision, 2018, pp. 805–822.

[49] X. Liang, K. Gong, X. Shen, L. Lin, Look into person: Joint body parsing & pose estimation network and a new benchmark, IEEE Transactions on Pattern Analysis Machine Intelligence 41 (4) (2019) 871–885.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[51] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Transactions on Image Processing 20 (8) (2011) 2378–2386.

[52] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444.

[53] R. Zhang, P. Isola, A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE Confefence on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[54] J. Wang, K. Sun, T. Cheng, et al., Deep high-resolution representation learning for visual recognition, IEEE Transactions on Pattern Analysis Machine Intelligence 43 (10) (2021) 3349–3364.